DOCUMENT RESUME

ED 393 890 TM 024 814

AUTHOR Edirisooriya, Gunapala
TITLE Stepwise Regression 1s a Problem, Not a Solution.
PUB DATE Nov 95
NOTE 16p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Biloxi, MS, November 8-10, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Employees; *Error of Measurement; Goodness of Fit; *Predictor Variables; *Regression (Statistics); *Research Methodology; Salaries; Sample Size
IDENTIFIERS *Stepwise Regression

ABSTRACT

Stepwise regression is not an adequate technique to provide the best set of variables with which to predict the dependent variable. By using the stepwise regression method, one who attempts to select the best set of predictors of a given dependent variable will face more problems than he or she attempted to resolve. This is illustrated with an example taken from the personnel data set of a medium-sized firm. An attempt was made to explain the variation in the present salary of the workforce in terms of a number of personnel variables. Stepwise regression techniques were applied for various sample sizes. The data are used to demonstrate that stepwise regression is quite vulnerable to specification, sampling, and measurement errors. It cannot be assumed that stepwise regression will provide the best-fit model, the order of significance of predictor variables, or the relative importance of predictor variables. Stepwise regression will not necessarily derive the appropriate solution for the researcher. (Contains two tables and nine references.) (SLD)

Stepwise Regression is a Problem, Not a Solution

Gunapala Edirisooriya

Department of Educational Leadership and Policy Analysis
East Tennessee State University

A Paper Presented at the Annual Meeting of the
Mid-South Educational Research Association
in Biloxi, MS, November 8-10, 1995

Stepwise Regression is a Problem, Not a Solution

Freed, Ryan, & Hess, (1991) in their *Handbook of Statistical Procedures and Their Applications to Education and the Behavioral Sciences*, commissioned by American Council on Education and published in the Macmillan Series on Higher Education, claim that the stepwise regression method is:

> A powerful technique commonly used to make predictions from several independent variables in a way that examines the power of the different independent variables to predict the dependent measure. (P. 65)

This is totally misleading. The Stepwise regression is not a powerful technique that can provide the best set of variables to predict the dependent variable. By using the Stepwise regression method, if one attempts to select the best set of predictors of a given dependent variable, then she/he will face more problems than she/he attempted to resolve.

The regression method can answer a question like the following: how much of the total variance of any given criterion variable (Y), can be explained by a given set of predictor variables ($x_1, \ldots, X_n$)? Algebraically,

$$Y = f(x_1, \ldots\ldots, X_n) \quad 1$$

How to decide on the appropriate set of predictors, $x_1, \ldots, X_n$, is a perplexing question. If a researcher follows the hypothetico-deductive paradigm in research, then the answer to this question can

be derived from the existing theory or theories. Therefore, what the researcher attempts to do is to test the existing theory. In contrast, a researcher who follows the inductive approach to research is trying to build a theory about the phenomenon being studied. In this case, this researcher doesn't know the exact set of predictors as in equation 1. This researcher may collect a larger set of variables, $g(x_1, \ldots, X_{n-1})$, where $g(x_1, \ldots, X_{n-1}) > f(x_1, \ldots, X_n)$. Now, the problem this researcher faces with is how to select the set $f(x_1, \ldots, X_n)$ out of $g(x_1, \ldots, X_{n-1})$. This is typically the case in social science research. Alternatively, a researcher may inadvertently select a subset, $h(x_1, \ldots, X_{n-1})$, of the appropriate set, $f(X_1, \ldots, X_n)$, as the initial data set for her/his analysis. The stepwise regression method, or any other method for that matter, is not going to find the appropriate solution of the set, $f(x_1, \ldots, X_n)$. Similarly, a researcher may start with a completely different data set and the chances of finding the appropriate solution is absolutely zero. No further elaboration is needed. We still have to prove that even if a researcher starts with a set of data, $g(x_1, \ldots, X_{n-1})$, of which $f(x_1, \ldots, X_n)$ is a subset, the stepwise regression method does not necessarily derive the appropriate solution for the researcher.

So, by using the Stepwise regression method, our attempt is to select the set of predictors, $f(x_1, \ldots, X_n)$, out of our larger set of variables $g(x_1, \ldots, X_{n-1})$. The Stepwise regression method searches

among the set of variables, $x_1, \ldots, X_n$, and find the predictor variable that produces the highest F value and proceeds, according to a set criteria, to enter variables (the remaining variables) and delete (the variables which were already included in the equation). The process of entering and deleting of variables ends when it meets the set criteria. The popular notions that (1) the subset of predictor variables selected from $g(x_1, \ldots, X_n)$ is the best set of variables in explaining the criterion variable Y and (2) the order of variable entry signifies the order of importance of x_i on Y have been dispelled by others (Thompson, in press & 1989; Snyder, 1991; Huberty, 1989; and Thompson, 1985).

Theoretically, when we submit this set of possible predictor variables, $g(x_1, \ldots, X_n)$, at every step of variable entry we assume that the coefficients of the variables not-in-the-equation are assumed to be zero. This raises the following issue: do the popular software packages correctly calculate applicable degrees of freedom for the numerator in calculating F statistics? This issue also has been highlighted by Thompson (in press) and Cliff (1987). Therefore, I concentrate on the following question: how robust is the stepwise regression method to sampling errors, measurement errors, and specification errors? If the Stepwise regression method is not robust to these errors, then the solution derived by this method does not necessarily guarantee to select the appropriate set, $f(x_1, \ldots, X_n)$,

out of the set, $g(x_1, \ldots, X_n)$. Then, the solution of the stepwise regression method is only *a* solution.

I analyzed a personnel data set of a medium-sized firm. The objective was to explain the variation in the present salary (SALNOW in dollars) of the workforce in terms of a number of personnel variables (x_i):

AGE = employee's age (in years)

EDLEVEL = educational level (years of education)

ETHGEND = educational level (0 = white male, 1 = non-white male, 2 = white female, 3 = non-white female)

ETHNIC = employee's ethnic background (0 = white, 1 = non-white)

GENDER = employee's gender (0 = male, 1 = female)

JOBCAT = employment category (1 = lowest, 7 = highest)

SALBEG = beginning salary (in dollars)

TIME = job seniority (in months)

WORK = work experience (in years)

The variable, ETHGEND, was created using GENDER and ETHNIC variables. Correlation coefficient matrix of the total data set revealed that the two variables, TIME (job seniority) and WORK (work experience), were highly correlated. The details are given in Table 1.

TABLE 1 HERE

Therefore, I experimented with two sets of predictor variables: one set which included the variable WORK, (regression equation mode denoted by **a**) and another set which excluded the variable WORK, (regression equation mode denoted by **b**). The data set included 474 cases with no missing data. To test the sampling robustness, I selected a number of **random** samples of sizes: 49, 85, 95, 111, 137, 145, 159, 186, 207, 239, 283, 331, 359, and 386. For each sample, I ran stepwise regression modes **a** and **b**, one set with standardized coefficients and another set with unstandardized coefficients. The results of the stepwise regressions with standardized and unstandardized coefficients were comparable. Therefore, and for the simplicity in interpretation as well, I opted for the regression equations with standardized coeffcients. The regression results are summarized in Table 2.

TABLE 2 HERE

The criteria set for PIN and POUT were 0.025 and 0.05 respectively and I used the *SPSS for Windows, Professional Statistics* for data analysis. As can be seen from Table 2, for 1) the various sample sizes and for 2) both the regression modes **a** and **b**, the stepwise regression method selected numerous combinations (in number and in composition) of predictors of the criterion variable, SALNOW. The number of predictors were two and six for the sample sizes of 49

(smallest) and 474 (largest) respectively.

For the sample sizes 49, 207, and 359 the Stepwise regression method selected the same predictor variables for both the regression modes ***a*** and ***b***. Generally, for the regression modes ***a*** and ***b***, when the number of selected predictors were the same but the compositions were different, the set of predictors for the regression mode ***a*** explained the variance in SALNOW marginally better. Not always. When the sample size was 159, the predictor set ***b*** gave a better fit in explaining the variation in SALNOW. When the sample sizes were 95, 111, 137, 145, 186, 207, 331, and 386 the number of predictors selected under the regression modes ***a*** and ***b*** were different. So, the predictors selected by the stepwise regression method is vulnerable to specification errors. A researcher who rely on the stepwise regression method to select the best-fit predictors has to find a way to make sure that she/he, in fact, is free of specification errors. This condition alone does not guarantee that the stepwise regression method will derive the best-fit model. A researcher can include the universal set of all possible predictors, $g(x_1, \ldots, X_n)$. Nevertheless, the stepwise regression method can still fail to select the best-fit model consisting of the subset of predictors, $f(x_1, \ldots, X_k)$.

For example, for the sample size 159, the stepwise regression mode ***a*** selected the following equation (8a): SALNOW = f(EDLEVEL,

JOBCAT, SALBEG, WORK) with an adjusted R of 0.97594. The stepwise regression mode ***b***, when the WORK variable was removed, selected the following equation (8b): SALNOW = f(AGE, JOBCAT, SALBEG, TIME) with an adjusted R of 0.97652. The equation 8b is marginally better than the equation (8a). Furthermore, in equation 8a, the first three predictor variables are highly correlated whereas, there are only two highly correlated predictor variables in equation 8b. As Johnston (1984) and Darlington (1968) have explained the notion of *independent contribution to variance* has no meaning when predictors are highly intercorrelated. The regression equation mode ***a*** did not derive the best solution (8b) when the variable WORK was included in the set $g(x_1, \ldots, X_{n-1})$, the universal set. Therefore, the stepwise regression procedure, starting with the predictor variable with highest F value does not necessarily guarantee to produce the best subset of predictors. Further evidence to the criticisms of Thompson, (in press), Snyder, 1991, and Huberty, 1989. Furthermore, as the regression equation solutions 8a and 8b show, a researcher can include the data set $g(x_1, \ldots, X_{n-1})$, the universal set, and may assume to be free of specification errors. Nevertheless, the stepwise regression method does not necessarily produce the best-fit solution. Therefore, for the researchers who heavily rely on the inductive approach to research, the stepwise regression method does not provide solid grounds to defend a model derived through this analytical technique. Furthermore, the stepwise regression method is

susceptible to sampling errors as well.

For the sample size 137, the Stepwise regression selected two predictors (SALBEG and WORK) for the regression mode ***a*** whereas for the same sample size, this method selected four predictors (AGE, EDLEVEL, JOBCAT, and SALBEG) for the regression mode ***b***. Basically, there was no difference between the two regression equation solutions in their explanatory power of the variance in SALNOW. In contrast, for the sample size of 145, the Stepwise regression selected four predictors (EDLEVEL, JOBCAT, SALBEG and WORK) for the regression mode ***a*** whereas for the same sample size, this method selected two predictors (JOBCAT and SALBEG) for the regression mode ***b***. Again, there was no difference between the two regression equation solutions in their explanatory power of the variance in the criterion variable. If the solution varies as the sample size varies, then a researcher cannot be so sure of the solution chosen for her\him by the stepwise regression method.

We also know that more predictor variables doesn't necessarily explain more variance in the criterion variable. For the sample size 49, the Stepwise regression method selected the smallest number of predictor variables SALNOW = f(EDLEVEL, SALBEG) with an adjusted R of 0.96. For the same sample size, SALNOW = f(SALBEG) alone will give an adjusted R of 0.94. As the sample size increased, the number

of predictors, in general, increased but the increase in the adjusted R was marginal. For example, as the sample size increased beyond 300, the number of predictors selected for the regression equation were either five or six between the regression modes ***a*** and ***b***; and the adjusted R^2 remained at a steady level of 0.96. Therefore, irrespective of the sample size, the addition of more predictor variables does not make a significant difference in the ability to explain the variance in the criterion variable. As much as these results show the vulnerability of the stepwise regression method to specification and sampling errors, these results also suggest that it would be inappropriate to use the stepwise regression method where the data are subjected to wide margins of measurement errors. Thompson (in press) and Huberty (1989) have discussed this aspect in detail.

As I have discussed, the stepwise regression method is quite vulnerable to specification, sampling, and measurement errors. Therefore, we have to question the belief or the general perception of researchers that the stepwise regression method provides a) the best-fit model, 2) the order of significance of predictor variables, and 3) the relative importance of predictor variables. A researcher with no theoretical or conceptual knowledge of the phenomenon being studied rests her/his faith in statistical procedures to provide that piece of information. The great danger in that practice is the

inherent possibility to derive theoretically and practically useless explanations of social phenomena (Thompson, in press; Snyder, 1991; and Cliff, 1987). Therefore, the stepwise regression method becomes a problem; not a solution in the absence of a well conceived theoretical or conceptual understanding of the issue being analyzed. One solution for those who advance data-driven theories is to come up with a computer program capable of calculating iterations of all possible combinations of predicto. variables to explain the variance in the criterion variable. Clearly, even if we assume that the model is specification free and the data are free of measurement errors, as the number of possible predictors increases, such a calculation procedure will be a monumental task. Alternatively, other variable reduction methods such as discriminant, principal component, or factor analyses can be utilized in explaining a given social phenomenon.

BIBLIOGRAPHY

Carver, R. P. (1978), The Case Against Statistical Significance Testing, in Harvard Educational review, vol. 48, pp. 378-399.

Johnston , R. B. (1968), Multiple Regression in Psychological Research and Practice, in Psychological Bulletin, vol. 69, pp. 161-182.

Freed M. N., J. M. Ryan & R. K. Hess (1991), Handbook of Statistical Procedures and Their Computer Applications to Education and the Behavioral Sciences, New York: American Council on Education and Macmillan Publishing Company.

Huberty, C. J. (1989), Problems With Stepwise Method--Better Alternatives, in B. Thompson (Ed.), (1991), Advances in Social Science Methodology, vol. 1, pp. 43-70, Greenwich, CT: JAI Press.

Johnson, J, (1984), Econometric Methods, (3rd ed.), New York: McGraw-Hill.

Snyder, P. (1991), Three Reasons Why Stepwise Regression Methods Should Not Be Used By Researchers, in B. Thompson (Ed.), (1991), Advances in Educational Research: Substantive Findings, Methodological developments, vol. 1, pp. 99-105, Greenwich, CT: JAI Press.

Thompson, B. (In press), Stepwise Regression and Stepwise Discriminant Analysis Nee Not Apply Here: a Guidelines editorial in Educational and Psychological Measurement.

Thompson, B. (1989), Why Won't Stepwise Methods Die? In Measurement and Evaluation in Counseling and Development, vol. 21, no. 4, pp. 146-148.

Wonnacott, R. J. & T. H. Wonnacott (1979), Econometrics, (2nd ed.), New York: John Wiley & Sons.

Table 1. Correlation Coefficient Matrix for n = 474

	AGE	EDLEVEL	ETHGEND	JOBCAT	ETHNIC	SALBEG	SALNOW	GENDER	TIME	WORK
AGE	1.00									
EDLEVEL	-0.28*	1.00								
ETHGEND	0.13*	-0.31*	1.00							
JOBCAT	-0.09	0.50*	-0.33*	1.00						
ETHNIC	0.11*	-0.13*	0.85*	-0.18*	1.00					
SALBEG	-0.01	0.63*	-0.38*	0.77*	-0.16*	1.00				
SALNOW	-0.15*	0.66*	-0.40*	0.76*	-0.18*	0.88*	1.00			
GENDER	0.05	-0.36*	0.47*	-0.32*	-0.08	-0.46*	-.045*	1.00		
TIME	0.05	0.05	0.01	-0.05	0.05	-0.02	0.08	-0.07	1.00	
WORK	0.80*	-0.25*	0.04	0.01	0.15*	0.05	-0.10	-0.17*	0.01	1.00

* = the coefficients which are statistically significant at an $\alpha = 0.025$

Table 2 Stepwise Regression Results of SALNOW on Two Sets of Predictor Variables

#	AGE	EDLEVEL	ETHGEND	JOBCAT	ETHNIC	SALBEG	GENDER	TIME	WORK	N	adj. R^2
1a						★				49	0.93869
1b						○					0.93869
2a		☆				☆				49	0.95931
2b		○				○					0.95931
3a						★			★	85	0.95361
3b	○					○					0.95234
4a				★		★			★	95	0.95917
4b				○		○					0.95712
5a		★		★		★			★	111	0.96156
5b		○		○		○					0.96000
6a						★			★	137	0.94963
6b	○	○		○		○					0.95332
7a		★		★		★			★	145	0.95441
7b				○		○					0.94988
8a		☆		☆		★			☆	159	0.97594
8b	○			○		○		○			0.97652
9a		☆		★		☆		★	☆	186	0.96742
9b	○			○		○		○			0.96666
10a	☆	☆		★		★				207	0.96799
10b	○	○		○		○					0.96799
11a				★		★	★	★	☆	239	0.96719
11b	○		○	○		○		○			0.96715
12a				★		★	★	★	★	283	0.96532
12b	○	○		○		○	○	○			0.96505
13a		☆		★		☆	★	★	★	331	0.96171
13b	○	○	○	○		○					0.96096
14a	★	☆	★	★		★		★		359	0.96458
14b	○	○	○	○		○		○			0.96458
15a				★		★	☆	☆	☆	386	0.96350
15b	○	○		○		○	○	○			0.96335
16a		★		★		★	★	★	★	474	0.96526
16b	○	○		○		○	○	○			0.96482

a. ★ = Predictors selected by stepwise regression including the variable, WORK

b. ○ = Predictors selected by stepwise regression excluding the variable, WORK